Exploring the Validity of a Comprehensive Listening Test to Identify Differences in Primary School Students' Listening Skills

Bourdeaud'hui Heleen[a], Aesaert Koen[b], Johan van Braak[a]

[a] Department of Educational Studies, Ghent University, Ghent, Belgium

[b] Ctr Educ Effectiveness & Evaluat, Katholieke Univ Leuven, Leuven, Belgium

Abstract

Effective listening comprehension skills are an important prerequisite for the academic success of primary school students. However, the assessment of listening skills in the instructional language appears to have received only scant attention in the literature. Therefore, the goal of the present study was twofold. Firstly, a comprehensive listening test was developed and different aspects of construct validity supporting the use of the listening test were explored. The listening test was administered to 1001 sixth-grade primary school students in Flanders, the Dutch-speaking part of Belgium. Next, the test items were controlled for item difficulty and discrimination, dimensionality, model-data fit, local item independence, monotonicity, and gender differential item functioning. The final listening test consisted of 24 multiple-choice and open-ended test items. Secondly, the listening test was used to identify differences between students' listening skills based on gender and home language. The results indicate that gender was not significantly related to listening comprehension skills, but L1 Dutch-speaking students significantly outperformed L2 Dutch-speaking students. This study also covers possible further fine-tuning of the instrument.

Introduction

Listening comprehension refers to the ability to process, integrate, and understand the meaning of spoken messages (Hogan et al., 2014). For primary school students, the ability to listen effectively in the instructional language is of critical importance for different reasons, such as acquiring and processing new information, understanding instructions from the teacher, participating in class or small group discussions, and developing other language skills (Acat et al., 2016; Adelmann, 2012; Andringa et al., 2012; Goh & Aryadoust, 2016; Iwankovitsch,

2001; Marx et al., 2017; Özbay, 2010; Wolfgramm et al., 2016). To determine students' listening skills in the instructional language and to understand the degree to which students master the listening curriculum, reliable listening instruments are essential (Acat et al., 2016; Buck, 2001; Rost, 2011). However, the extent to which existing listening tests can reliably measure the listening construct can be disputed (Santos et al., 2015). Listening tests are frequently included in reading batteries or inventories, in which a text is read aloud to the students, and they have to read and answer some questions on the test paper. These listening test results may depend largely on students' reading and writing abilities instead of capturing the listening construct (Acat et al., 2016; Brownell, 2016; Green, 2017; Özbay, 2010; Rost, 2011; Santos et al., 2015).

In developing and evaluating tests, validity is a fundamental consideration and consists of the collection of various empirical evidence "for or against the defensibility of inferences drawn from test scores" (Roever & McNamara, 2006, p. 234). While increasing attention has been paid to the construction and validation of second language (L2) listening tests, the interest for assessing listening skills in the instructional language remained far behind (Buck, 2001; Flowerdew & Miller, 2010; Rost, 2011). This lack of focus could be due to the long-standing predominant assumption that students already have a good level of listening skills when they enter primary school (Lau, 2017). Indeed, it is clear that most L1 listeners have a large advantage compared to L2 listeners, as they can decode the oral input automatically without spending time and energy on translating words and phrases (Brown, 2008; Siegel, 2013). However, not all primary school students are naturally good at listening in the instructional language (Brown, 2008). In Flanders (the Dutch-speaking part of Belgium), large-scale assessments showed that one in five students did not reach the attainment targets for listening at the end of primary school (Authors, XXXX).

Since validity belongs to interpretations of test scores and not to tests themselves, validation efforts should be concentrated on the uses of test scores (Zumbo & Chan, 2014). Language tests are often used to make decisions about outcomes for individual students or compare different groups of students (Bachman & Palmer, 2010). For example, a trend in the listening literature has been to identify listening differences between boys and girls. However, until now, the results are rather inconclusive: whereas some studies has indicated that gender was positively related to students' listening skills in favor of girls (Oduolowu & Oluwakemi, 2014), this relationship could not be confirmed in other studies (Lin et al., 2015; Wolfgramm et al., 2016). Further, a vast amount of research has identified differences between native and non-native listeners, showing that the latter group was more likely to have lower levels of listening

ability (e.g., Oduolowu & Oluwakemi, 2014; Wolfgramm et al., 2016). In Flanders, over 20% of the students do not speak the instructional language at home and might be at higher risk for listening difficulties (Pulinx & Van Avermaet, 2014). More research is necessary to provide a comprehensive view of the relationship between home language, gender, and listening comprehension skills.

In summary, this study aimed to contribute to the unexplored field of listening assessment in the language of schooling by developing a comprehensive listening test and exploring different aspects of validity evidence in order to provide a high-quality measurement instrument that can be used to identify differences in students' listening skills. After briefly outlining the theoretical construct of listening, the focus was placed on content and internal validity aspects as put forth in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA et al., 2014]; further referred to as the *Standards*).

Literature Review

The Standards for Educational and Psychological Testing

Validity is not an inherent property of the test, but the degree to which we can justify the test's score-based inferences (Messick, 1989). Construct validity refers to the concept that a test is designed to measure and is often considered as "the overarching validity concept" (Chapelle, 1999, p. 257). Several frameworks have been proposed to explore the construct validity of educational tests (e.g., AERA et al., 2014; Bachman & Palmer, 2010; Kane, 2013; Messick, 1996). *The Standards* (AERA et al., 2014) listed five possible sources of validity evidence that test developers can use to increase the quality of their assessment instrument, i.e., evidence based on (a) test content, (b) internal structure, (c) testing consequences, (d) relations to other variables, and (e) response processes.

First, validity evidence based on *test content* concerns making various decisions about the way the listening construct is measured, such as the choice of the text topic, item format, item wording, and guidelines for test administration and scoring. Second, validity evidence based on the *internal structure* of a test is defined as the degree to which test items conform to the latent construct. Third, evidence about *relations to other variables* concerns the correlation

between test scores and variables external to the instrument, for example, by convergence with other tests. Fourth, *consequential validity evidence* refers to the effects of the interpretation and decisions based on the test scores. Finally, validity evidence-based on *response processes* can be gathered by using interviews, think-aloud procedures, or surveys to question test takers about their answers on the test (AERA et al., 2014; Zumbo & Chan, 2014). This study mainly focused on validity evidence based on *test content* and *internal structure* for the listening instrument.

Validity Evidence based on Test Content

Content-oriented evidence refers to the analysis of the relationship between the test content and the construct it is supposed to measure. Validity evidence based on test content can be observed in, for example, the themes, wording, format and scoring of test items, and test administration (AERA et al., 2014). Two important threats to test content validity, which may give an unfair (dis)advantage to specific subgroups of test takers, are construct-irrelevant variance and construct underrepresentation (AERA et al., 2014; Messick, 1996). In this respect, some important issues may arise in the development of listening tests.

First, in most listening tests, students have to read the test items in silence after listening to the stimulus text. However, language researchers agree that only providing the test items in written mode on the test paper can undermine the reliability of the listening test because a construct-irrelevant factor, i.e., reading ability, is measured (Chang & Read, 2013; Yanagawa & Green, 2008). Lower-level reading students may be disadvantaged or even give up reading the test items and just guess the answer. The written mode can also distract test takers' attention on the listening input, as they have to switch quickly from listening to reading ability (Chang & Read, 2013; Yanagawa & Green, 2008). On the other hand, presenting test items only in the spoken form makes the test a purer measure of listening but requires a good working memory capacity, again threatening the content-oriented validity of the listening test (Chang & Read, 2013; Weir, 2005). The spoken form may also increase test takers' anxiety and provoke guessing as students can forget the test items (Buck, 2001; Chang & Read, 2013). Considering the disadvantages of presenting listening test items in a single written or spoken mode, it is recommended to present test items in both spoken and written mode to the students.

A second content-related issue in the development of listening tests is related to the presentation of the stimulus text. Traditionally, teachers read stories aloud to the class during listening assessment. When audio technology was introduced, audio recordings were played to the students, in which students only heard the spoken words. Later, the spreading of video

technology made it possible to use video fragments during listening assessment, involving both aural and visual stimuli. Language researchers agree that videos reflect a higher level of authenticity and might lead to a more construct-relevant assessment of listening skills. The dual-channel representation of sound and vision replicates a real-life listening event more closely, as in most listening situations the listener is able to see the speaker, and oral information is accompanied by visual information (Buck, 2001; Field, 2013; Ginther, 2002; Ockey, 2007; Sulaiman et al., 2017; Wagner, 2013). Additionally, as listeners may vary in their ability to interpret and utilize the non-verbal information provided by the speaker, not including the visual channel in listening tests may lead to construct underrepresentation. In this respect, the listening test may fail to capture important aspects of the listening construct, such as understanding the nonverbal components of spoken communication (e.g., body language and facial expressions) (Buck, 2001; Ockey, 2007; Wagner, 2008). Despite the importance of visuals, its use in listening assessment in the language of instruction has been rather limited so far.

Further, evidence based on test content can be found in the correspondence between the content and students' listening standards or in expert judgments of the appropriateness of the measurement instrument (AERA et al., 2014). These concerns will be discussed later in the test development process.

Validity Evidence based on Internal Structure

Validity evidence based on internal structure provides information on how relationships among test items and components are true to the construct that the test intended to measure (AERA et al., 2014). Traditional classical test theories and/or more modern measurement methods are often used to assess a test's internal structure (Rios & Wells, 2014). In classical test theories (CTT), the focus is at the level of test scores. Therefore, the test taker and item characteristics are interdependent and can only be understood in the context of each other (Crocker & Algina, 2006; Hambleton et al., 2013). Modern test theories, such as item response theory (IRT), can overcome these limitations as they examine responses at the item level (Embretson & Reise, 2013). As most test development projects today rely on CTT and IRT (Davidson, 2004), both types of analyses will be used in this study.

Dimensionality, reliability, and item invariance of the measure are three fundamental aspects to provide evidence of internal structure validity (Rios & Wells, 2014; Rupp & Leighton, 2016). First, *the dimensionality* of a scale determines how many latent variables are assessed, i.e., if the measure is unidimensional or multidimensional. A test that intends to report

one composite score, such as listening comprehension, should be predominantly unidimensional. Model-based approaches, such as IRT can examine whether item responses fit the unidimensionality. Second, *reliability* implies that the measurement instrument produce internally consistent outcomes. The CTT approach assesses internal consistency with the statistic Cronbach's alpha, whereas within the IRT framework, the item parameter estimates are used to quantify reliability (Rupp & Leighton, 2016). Third, *item invariance* indicates whether particular items function differently for some subgroups. An item is functioning differently when different groups of test takers with similar overall ability have systematically different probabilities of answering the item correctly. Differential item functioning (DIF) is a commonly used method to judge whether the test items are fair and free from systematic bias (Ferne & Rupp, 2007; Magis et al., 2010; Song et al., 2015). Over recent decades, DIF analyses have been conducted with various language tests, but to our knowledge, they have not yet been conducted with listening tests in the language of instruction.

Listening Differences between Student Groups

A vast amount of research has focused on gender differences in listening skills in the language of instruction. It could be expected that girls are better listeners as they were found to outperform boys in other language skills such as reading (Lynn & Mikk, 2009), have an advantage in vocabulary tests (Asia et al., 2019), and seem to have fewer concentration problems (Wolfgramm et al., 2016). However, research investigating the relationship between gender and listening skills has produced mixed findings. Whereas some studies indicated that gender is positively related to students' listening skills in favor of girls (e.g., Oduolowu & Oluwakemi, 2014), other studies found no significant gender differences (e.g., Lin et al., 2015), or even found a small advantage in favor of boys (e.g., Wolfgramm et al., 2016).

A possible explanation could be that - next to students' listening ability - some items of the listening test measured an unknown characteristic that is more present in girls than in boys or vice versa. Multiple studies have shown that variation in item characteristics (e.g., passage topic, item location, content, and vocabulary) is sensitive to gender differences and may cause gender-based DIF (Aryadoust et al., 2011). Over recent decades, DIF analyses have been conducted with various language tests, but much less attention has been paid to gender differences on standardized listening tests in L1 contexts. As such, it is not clear whether statistical differences between boys and girls reflect true differences in listening skills or are

due to test characteristics (Oliveri et al., 2013). Taking DIF into account would make it possible to identify non-biased differences in listening ability between boys and girls.

Next to gender, differences between students' listening skills are often studied with a focus on home language. Various studies showed that primary school students who do not speak the language of schooling as their mother tongue experienced a greater challenge for developing listening skills in comparison to their native-speaking classmates (Andringa et al., 2012; McKendry & Murphy, 2011; Oduolowu & Oluwakemi, 2014; Wolfgramm et al., 2016). In this study, the comparison between native and non-native students was excluded from DIF analysis because these are not different groups with the same ability level but rather groups with different expected ability levels.

In Flanders, there is a strong monolingual education policy, considering the language of instruction as the only legitimate language and minority languages rather as a barrier for academic success (Agirdag, 2010; Pulinx & Van Avermaet, 2014). Besides, minority languages are considered as more or less valuable based on their attributed social status (Blommaert & Van Avermaet, 2008; Bourdieu, 1991). In the context of migration, integration, and citizenship, Western European languages such as French, English, German, and Chinese are classified as high-status languages, whereas Eastern languages such as Turkish, Moroccan, and Arabic are considered as low-status languages. In Flemish educational policies, there is the implicit preconception that students who have a low-status home language are at higher risk for school failure (Pulinx & Van Avermaet, 2014). This predominant policy may impact the general beliefs teachers hold and may also influence students' self-concept, motivation, and learning opportunities, all of which may affect student literacy outcomes (Pulinx & Van Avermaet, 2014; Pulinx et al., 2017).

Research Aims

The lack of focus on the assessment of listening tests in the language of schooling and the importance of drawing comparisons across respondent groups provided the rationale for this study. This research was organized into two partial studies. The first part aimed to explore the validity of a comprehensive listening test for primary school students in Flanders. The listening test was developed for sixth-grade students as upper-primary school students have been highly under-represented in listening research, and most listening research focused on preschool or lower primary school students (Beall et al., 2008). The second part of this study aimed to

identify the relationship between home language, gender, and students' listening skills in the language of schooling using the developed measurement instrument. With regard to the second part, the following two research questions were addressed:

(1) Are there differences between boys and girls in listening skills as measured by the listening instrument?
(2) Are there differences between native and non-native Dutch-speaking students' listening skills as measured by the listening instrument?

Test Development

The *Standards* recommend that the test development process should be guided by a set of test specifications, including the description of the target construct, design aspects of the test, format/scoring of the test items, test administration, as well as the psychometric specifications to analyze the statistical properties of the items and the whole test (AERA et al., 2014).

Content Specifications

A first step in the listening test development process was to clearly define the construct of listening skills. Based on the listening literature, listening skills can be operationalized into two types of comprehension: (a) literal comprehension, the understanding of information explicitly stated in the text and (b) inferential comprehension, the understanding of implicit information (Brownell, 2016; Karimi & Naghdivand, 2017; Kim, 2015; Potocki et al., 2012; Santos et al., 2015). Literal comprehension refers to the ability to recall information directly presented in the text, such as details and facts, and implies that the listener decodes and analyzes the auditory message (Brownell, 2016; Santos et al., 2015). Inferential comprehension refers to the ability to combine visual, auditory, and situational information to fully understand a message that is not explicitly mentioned in the text (Brownell, 2016; Potocki et al., 2012; Santos et al., 2015). Inferential listening requires the use of skills such as making inferences between different text parts, predicting outcomes, determining why an event is told, and finding the main idea (Potocki et al., 2012).

In a second step, a literature review was conducted to translate these two general components into measurable sub-skills. Literal comprehension could be further subdivided into three sub-skills: (1) defining the literal meaning of a word or a word group, (2) identifying information that has been explicitly mentioned in the stimulus text, and (3) remembering and

identifying facts and details from the stimulus text (e.g., time, place) (e.g., Brownell, 2016; Karimi & Naghdivand, 2017). Inferential comprehension could also be further subdivided into three sub-skills: (1) deriving the implicit meaning of a word or a word group, (2) identifying simple or complex relationships between sentences or larger text parts (e.g., cause and effect relationship), and (3) identifying the global content of the stimulus text (e.g., Brownell, 2016; Karimi & Naghdivand, 2017). These skills were integrated into a multilayered test framework, which was considered as the heart of the test development process and acted like a theoretical blueprint for the development of the listening test items.

In total, 46 test items were developed according to the skills of the test framework, i.e., 24 items measured students' literal comprehension skills, and 22 items measured their inferential comprehension skills. Table 1 presents the final test framework with the different skills and some example items.

An independent and diverse panel of experts consisting of two sixth-grade teachers, three educational advisers experienced in listening skills, and two test developers provided feedback on the development of the test framework and the test items under construction. More specifically, the panel of experts criticized the test items by reviewing test content for language, illustrations, and other representations that might be interpreted differently by different student groups, pointing to potential sources of irrelevant variance. Further, the experts judged the degree to which the item content matched the content categories of the test framework and the listening curriculum, and whether the listening test provided balanced coverage of the listening construct. According to their feedback, different test items were reformulated or replaced by new items.

As the listening test would be administered to sixth-grade students in Flanders, the test must take the specific content of the Flemish listening curriculum into account. This listening curriculum included ten attainment targets or minimum objectives for all students to master by the end of primary school. Considering this curriculum as a blueprint for acquiring effective listening skills in the class context, two out of ten attainment targets were selected for the listening test, i.e., (1) students must process information from *an informative text* and (2) students must process information from *a teacher instruction*. These attainment targets were selected for the listening test as they are frequently addressed in the primary school context and we expected that students were familiar with them. Consistent with the selected attainment targets, two text types had been put forward, i.e., (1) an informative text and (2) a teacher instruction.

Test Design Specifications

To minimize construct-irrelevant variance and to promote valid score interpretations for the intended use of the test, a number of well-balanced choices about test design aspects were made. First, the listening test was offered in the format of a *video file*, simultaneously offering aural and visual input, because videos more closely simulate real-life listening situations and might reflect a greater level of listening authenticity (Wagner, 2013). More specifically, (1) six informative texts (four short and two long texts), selected from the daily Dutch youth news program *Karrewiet* and (2) two recorded instructions with practical assignments from a fictional teacher – recorded by a native Dutch female speaker, were administered to the students. For the instructions, the topics *a trip to a museum* and *a trip to an animal park* were chosen as general themes, as these were not explicitly part of the course content but are still recognizable enough for the target population. In addition, subjects, words and expressions that were especially associated with specific cultural backgrounds, socioeconomic status, or ethnic groups, were avoided to minimize confounding of this measurement with prior knowledge and experiences that are likely to (dis)advantage students from particular subgroups (AERA et al., 2014; McKendry & Murphy, 2011). Additionally, to better replicate real-world listening situations, every video was played only once to the students (Green, 2017). Figure 1 shows a screenshot of an informative text and a teacher instruction.

An additional concern was *the presentation mode of the test items*. To prevent that students' listening outcomes were largely influenced by their reading comprehension skills, every question (including both item stems and answering options) has been read aloud on the recording. The questions were also presented in a paper test booklet to avoid the influence of the working memory capacity.

Finally, note-taking was not allowed during the listening test, as this depends largely on the working memory buffer, and it can be challenging for young students to simultaneously take notes and focus on the continuous flow of incoming new information. Note-taking is comprised of a complex array of skills such as eye-hand coordination and writing skills that have been found to vary widely among students. Students who lack note-taking skills can feel overwhelmed, spending more time writing down notes and missing a big part of the text content (Piolat et al., 2005).

Item Format and Scoring Specifications

Multiple-choice questions are more objective, less time consuming, and less dependent on writing competences than open-ended questions (Buck, 2001; Green, 2017). This issue is especially pertinent when developing a listening test for non-native speakers, as their articulation proficiency in the language of instruction may be an extra challenge in open-ended questions (McKendry & Murphy, 2011). Therefore, the largest proportion of test items was developed with a multiple-choice format. However, a certain amount of open-ended test items were also integrated into the test to avoid exclusively testing recognition knowledge and inviting pure guessing (Buck, 2001). In total, 37 multiple-choice items with four answering options and 9 open-ended test items were developed. The multiple-choice items were scored dichotomously (1 = correct; 0 = incorrect), and a scoring key was developed to guide the assessment of the open-ended items (1 = correct; 0 = incorrect). Every question was read aloud on the recording and followed by a beep sound. After the beep sound, students had twelve seconds to answer a multiple-choice question and up to one minute to answer an open question. Students were not allowed to preview the questions before they watched the video.

Pilot Study

A pilot study was conducted to evaluate the usability of the test items under development. Therefore, the listening test was administered to three classes of sixth-graders (n = 42). During test observation, we primarily focused on (1) the quality of the video recording, (2) the difficulty level of the test items, (3) the comprehensibility of the test items and item instructions, and (4) the answering time. Regarding the difficulty level of the items, the results of the pilot study showed that four items were answered correctly by more than 95% of the students, and one item was answered correctly by only 2% of the students. These five items were adapted to decrease or increase their difficulty level. Besides, words in the stem questions or answering options that turned out to be too difficult or unclear for the target group were replaced by more appropriate vocabulary. Further, more information was collected about the time that students needed to complete each item. In this way, it was decided to provide less answering time on the recorded videotape for the multiple-choice questions (twelve seconds instead of twenty seconds) but more answering time for the open-ended questions (approximately one minute per question). Finally, some adaptations were made related to the test instructions, for example, the page numbers of the listening test were read-aloud on the

recorded videotape to prevent students from running through the pages and reading the next questions in advance.

Psychometric Specifications

Psychometric specifications refer to the desired statistical properties of the test items (e.g., item difficulty and discrimination) and the whole test (e.g., test difficulty and reliability) (AERA et al., 2014). In this study, CTT and IRT were considered as complementary approaches that provide useful information at various phases of the examination of the psychometric quality of the listening test (De Champlain, 2009). Item statistics based on CTT were helpful to identify weaknesses in the early phases of processing, whereas IRT was applied to estimate final item difficulties and item discrimination. Below, we outline the steps taken to examine the psychometric quality of the listening test.

In the first phase, the principles of classical item analysis were followed to calculate item difficulty index (i.e., the proportion of students that answered the item correctly) and item discrimination index (i.e., the correlation value between the score on the particular item and the total test score) (Zubairi & Kassim, 2006). According to item difficulty, a p-value of zero refers to a very difficult item, whereas a p-value of one indicates an item that is answered correctly by all respondents. In this way, we considered a p-value between .30 and .90 as desirable (Haladyna et al., 2002). Concerning item discrimination, good items should have a point-biserial correlation of or above .25 (Spaan, 2007).

In the second phase, it was examined whether the listening test could be perceived as a unidimensional or a multidimensional construct. Because the data were dichotomous, dimensionality was investigated using nonlinear exploratory factor analysis (NLFA) (de Ayala, 2009). Model fit statistics including Tanaka's (1993) Goodness of Fit Index (GFI) and the Root Mean Square Residual (RMSR) were used to determine which dimensional solution was the best. In general, a good model fit is indicated by a GFI-value over .90 and a small value of RMSR (de Ayala, 2009). RMSR-values equal to or smaller than four times the reciprocal of the square root of the sample size indicate a good model fit (de Ayala, 2009). Test dimensionality was defined as the model with the highest number of dimensions that still produces a 10% or greater decrease in the RMSR over the preceding model (Tate, 2003).

A basic assumption in the application of IRT is that the model fits the data (Edelen & Reeve, 2007). As such, in the third phase, the model-data fit was investigated through the comparison of model predictions and the observed data. For dichotomous items – as used in

this study – the Rasch and two-parameter logistic models (2PLM) are most commonly used. The Rasch model estimates only the item difficulty holding the item discrimination constant, while the 2PLM estimates both item difficulty and item discrimination. Absolute model-data fit was investigated using the Standardized Root Mean Square Residual (SRMSR) and MADaQ3 effect size for model fit statistics (Maydeu-Olivares, 2013). The closer the values of the SRMSR and MADaQ3 are to zero, the better the model fits the data. Relative model-data fit was investigated comparing the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for both models. Regarding AIC and BIC, lower coefficients are more desirable (Neumann et al., 2011). Besides, item infit statistics (i.e., the weighted mean-squared residuals between what is observed and what would be expected) were calculated, with infit statistics between 0.80 and 1.20 indicating good item fit (de Ayala, 2009).

Fourth, the assumption of local item independence (LII) was checked. LII means that responses to two different items should be statistically independent of each other for individuals at the same ability level (Hambleton et al., 1991). Yen's Q3 (1993) was used for the identification of local item independence. Commonly, .20 cut-points are used to identify items that violate the assumption of local independence (Yen, 1993).

Fifth, the assumption of monotonicity was checked. The differences between the actual and predicted performances were checked by comparing the item characteristic curves (ICC) with the plots of the observed values of each item (Hambleton et al., 1991; Yen & Fitzpatrick, 2006). When an item verifies the assumption of monotonicity, it is assumed that with an increase in ability, the probability of getting a correct response increases too (Reckase, 1997).

In the last phase, the Mantel-Haenszel (MH) method was used to identify items that display DIF for gender. The MH-method is a non-parametric chi-square test (Roussos et al., 1999) that works with item responses from two groups, referred to as the reference group (boys) and the focal group (girls). The following criteria were used to identify DIF-items: items which display negligible DIF are identified by $|\Delta_{MH}| < 1.00$; items which display moderate DIF are identified by $1.00 \leq |\Delta_{MH}| < 1.50$; and items which displayed large DIF are identified by $|\Delta_{MH}| \geq 1.50$ (Zieky, 1993). The items showing no DIF are selected as anchor items. Once the DIF-items were identified, a multiple group analysis in which the item parameters of the DIF-items are allowed to vary between the two groups, i.e., boys and girls, was conducted to estimate the final model.

After the examination of the psychometric quality of the listening test items, the empirical reliability index of the overall test was calculated. The range of reliability measures

was valued as follows: a) less than .50 show low, b) between .50 and .80 show moderate, and c) greater than .80 show high empirical reliability of the listening test (Tucker, 2007).

Participants

The psychometric specifications of the listening instrument were assessed through a large group of students. Therefore, the 46 items were administered to a representative sample of 1001 sixth-grade primary-school students in Flanders. Schools were stratified for school size (i.e., small school < 180 students; large school ≥ 180 students), province, and educational network (i.e., official public education, subsidized public-authority education, and subsidized private-authority education). Data were collected in 74 classes from March till May 2018. The mean age of the students was 11.88 years, with a minimum of 10.57 and a maximum of 14.54. Of the students, 50.7% (n = 508) were boys and 49.3% (n = 493) were girls.

The sample consisted of 829 native Dutch-speaking students (82.8%) and 172 students with a different mother tongue (17.2%). In total, 34 different first languages were represented, such as French (n = 59), Arabic (n = 20), Turkish (n = 14), Polish (n = 10), Berbers (n = 9), English (n = 8), Russian (n = 4), and Greek (n = 4). Of the students with a different home language, Western European languages such as French, English, and German were classified as high-status (n = 78), whereas Eastern languages such as Turkish and Arabic were classified as low-status (n = 94) (Pulinx & Van Avermaet, 2014). Appendix B gives an overview of the different represented languages and the classification in high- and low-status languages in Flanders.

Results

Psychometric Properties of the Listening Test

Classical Item Analysis

With regard to item difficulty, the p-values of five items were located outside the critical range of .30 and .90, indicating that these items were too easy or too difficult for the students and should be retained for further analysis. Item discriminations in the CTT paradigm were calculated via the item-total point biserial correlations. The item-total correlation of 14 items was located outside the range of .25. As these items did not sufficiently discriminate

between students, they were removed for further analysis, resulting in 27 items. Table 2 presents the item characteristics from the classical item analysis. Because 19 items have been deleted, an additional check-up was done to control if the different skills from the test framework were still covered by the items of the listening test.

Dimensionality

Further, a Nonlinear Factor Analysis (NLFA) was conducted on the remaining 27 items. A unidimensional and a two-dimensional solution were directed using the NOHARM function in R (sirt-package). To determine which dimensional solution was the 'best', the differences among the models were examined. The first analysis for obtaining the unidimensional solution showed that Tanaka's GFI had a high value of .99990 and the value of the RMSR (= .00554) was smaller than the critical value of .12643 [4*(1/√1001)]. The subsequent NOHARM analysis for obtaining the two-dimensional solution revealed a Tanaka's GFI of .99991, while the RMSR decreased to .00513. Based on the economic principle, the value of the RMSR must decrease below .00470 to confirm the two-dimensional solution. In this case, the GFI and the RMSR did not provide decisive evidence for a two-factor solution. Further, the loadings of the two-factor analysis showed that many items had low loadings on the second factor including many cross-loadings. As such, the NLFA provided sufficient support that the unidimensional model was an accurate representation of the data to proceed with the IRT calibration. Table 3 shows the factor loadings of the unidimensional solution. Due to low factor loadings ($\lambda < .300$), three items (items 17, 28, and 44) had to be removed. A second unidimensional solution with the remaining 24 items resulted in a GFI of .99992 and RMSR of .00506, indicating a good model fit.

Model-Data Fit

The SRMSR and MADaQ3 were close to zero for both the Rasch model and the 2PLM, indicating a good model-data fit for both models. Further study of the absolute and relative model fit indices showed that the AIC decreased, while the BIC increased a little between the Rasch and the 2PLM (Table 4). The Chi-square test was significant ($p < .001$), showing that the 2PLM was preferable over the Rasch model. Further, fit indices were calculated for each item. All item infit statistics lied within the critical range of 0.800 and 1.200 for both the Rasch (min = 0.954; max = 1.045; min *pholm* = .379) and 2PLM (min = 0.991; max = 1.019; min *pholm* = .639) (Table 5). However, the item infit statistics were closer to one for the 2PLM, confirming the use of the 2PLM.

Local Item Independence

The results showed that none of the items had a Q3-value higher than .20, indicating that they were not interrelated with other items and all items were considered as locally independent. As a consequence, test taker's responses to different items were statistically independent after taking the latent trait into account.

Monotonicity

Finally, the study of the item characteristic curves of the 24 remaining items indicated that none of the items violated the assumption of monotonicity. For all items, the chance of answering the items correctly increased as the students' ability level increased. Figure 2 illustrates for item 17 that the probability of a correct answer was increasing with the ability level.

Reliability

Finally, the empirical reliability of the overall test with the remaining 24 items was calculated. The value of *r* was .67, indicating that the scale developed for the measurement of students' comprehensive listening skills showed moderate to good internal consistency.

Differential Item Functioning

DIF-statistics were calculated using the MH-method. First, it was determined whether the listening test items functioned differentially for boys (reference group) and girls (focal group). Of the 24 listening items, seven items displayed DIF for gender. The results showed that only two items favored girls and five items favored boys. However, the two large DIF-items were in favor of girls, while the five moderate DIF-items favored boys. Table 6 displays the effect sizes and the corresponding p-values of the test items.

Once the DIF-items were identified, a multiple group analysis was conducted to re-estimate the 2PLM mirt package in R (Chalmers, 2012). This means that the parameters of the DIF-items were allowed to vary freely between the two groups, i.e., boys and girls.

Item Ability Scale

The remaining 24 items and their parameters under the 2PLM were renumbered and ordered by item difficulty on the item ability scale in Table 7. The items distributed at the top of the scale were categorized as the most difficult items, while the items distributed at the bottom of the scale were rated as the easiest items. In general, the results showed that items measuring literal or inferential listening skills were scattered on the ability scale, representing different difficulty parameters. For example, both the easiest and the most difficult item belonged to literal comprehension. With regard to item type, the easiest item was a multiple-choice question, while the most difficult item was an open-ended question.

Listening Differences between Student Groups

Finally, listening differences between the subgroups were compared based on the standardized listening test (min. score = -2.958, max. score = 2.973). A univariate GLM (ANOVA) was conducted to check for the main effects and interaction effects of gender and home language on students' listening skills in the language of schooling. The results are presented in Table 8. The findings indicate that the mean test score for native Dutch-speaking students (n = 829) was statistically significantly higher than the mean test score for non-native Dutch-speaking students (n = 172), F (1,1001) = 15.713, (p < .001). Further, the differences in test scores between girls (n = 493) and boys (n = 508) were not statistically significant (p > .05). Finally, the interaction between gender and home language was not statistically significant (p > .05), indicating that the differences between native and non-native Dutch-speaking students did not depend on gender.

To identify differences between high- and low-status language groups, a post-hoc comparison was conducted using the Games-Howell test. The results in Table 9 show that the listening skills of students speaking a high-status mother tongue (n = 78) did not statistically significantly differ from students with a mother tongue that was classified as low-status (n = 94).

Discussion

Effective listening skills in the instructional language are an essential prerequisite for primary school students' academic success (Acat et al., 2016; Adelmann, 2012; Andringa et al., 2012; Goh & Aryadoust, 2016; Iwankovitsch, 2001; Özbay, 2010; Wolfgramm et al., 2016;

Wolvin, 2012). However, the assessment of listening skills in the language of schooling received only scant attention in research and practice. This exploratory study responds to the call for more listening assessment research by introducing and exploring the validity of a comprehensive listening test. To guide this investigation, the focus was placed on validity evidence about the internal structure and the test content, as put forth in the *Standards* (AERA et al., 2014). Afterward, the listening test was used to identify listening differences between different student groups in Flanders.

First, this study aimed to collect evidence of validity to support the use of the developed instrument. Some well-grounded decisions about the design aspects of the listening test were made, which - along with the content analysis, linkage to the curriculum, expert panel, and pilot study - contributed to the evidence based on test content. More specifically, every test item (including both item stems and answering options) was presented in written and spoken form to prevent that students' listening outcomes were largely influenced by their reading comprehension skills or working memory capacity. Further, the listening test consisted of video files instead of audio-only fragments, as excluding the visual channel from listening tests may lead to construct under-representation.

Besides, we collected evidence of *internal validity* that favors the use of the listening instrument. More specifically, we examined the psychometric quality of the listening test items through classical test theory statistics, item response theory analysis, and differential item functioning. Factor analysis revealed that the instrument measured a single latent trait, labeled as students' comprehensive listening ability. This result is in line with earlier listening research delivering theoretical and empirical evidence for listening comprehension as a single or unidimensional construct (Berninger & Abbott, 2010; Lehto & Antilla, 2003). The results also showed that the 2PL model, in which the items had unique discrimination and difficulty parameters, was a good fit to the data. With regard to the distribution of the items along the ability scale, the findings indicated that the literal and inferential items were scattered on the ability scale. These overlapping item difficulties showed that the relationship among the literal and inferential levels was not hierarchical, and literal items could be either easier or more difficult than inferential items. Finally, it was determined whether the items worked differently when processed by boys or girls. The results showed that approximately one-third of the listening test items were identified as gender DIF-items.

In particular, boys and girls displayed different probabilities of successfully completing test items with specific linguistic elements. It is possible that these items showed a larger magnitude of DIF because the students had to use their prior vocabulary knowledge to

determine the meaning of the requested words. Therefore, linguistic items in listening tests should be largely context-dependent, which means that students can determine the meaning of the requested word based on their comprehension of textual information and not by using prior experiences and knowledge (Jang & Roussos, 2009). To better understand the factors that affect DIF in listening tests, future research should study more closely the contribution of test item content, but also the influence of other test characteristics, such as item format, stem length, and the number and attractiveness of distracters (Aryadoust et al., 2011).

The second aim of this study was to examine whether differences in comprehensive listening skills were related to gender and home language. Concerning the relationship with gender, the results showed that, after controlling for gender DIF, girls and boys did not score significantly different for listening skills, adding new evidence to conflicting findings (e.g., Lin et al., 2015; Oduolowu & Oluwakemi, 2014; Wolfgramm et al., 2016). In this respect, listening skills differ from other language skills, such as reading skills where girls mostly significantly outperform boys (e.g., Lynn & Mikk, 2009).

Further, the results indicated that native Dutch-speaking students scored significantly higher than their non-native speaking peers. These results are comparable to earlier listening research showing that non-native speaking students have lower listening performance in comparison to their native-speaking peers (e.g., Andringa et al., 2012; McKendry & Murphy, 2011; Oduolowu & Oluwakemi, 2014; Wolfgramm et al., 2016). Different factors can account for this gap between native and non-native speakers. A higher proportion of unknown words may cause lexical gaps and interrupt the continuous process of listening (Hagtvet, 2003). Additionally, non-native listeners are often less capable of processing prosodic information in the language of schooling (Akker & Cutler, 2003; Andringa et al., 2012; Papadopoulou & Clahsen, 2003), or have a lower self-concept and confidence, which may impede their listening skills (e.g., Serraj & Noordin, 2013). Finally, no differences were found between students' speaking a high-status home language (such as English, French, German, and Chinese) and students' speaking a low-status home language (primarily Turkish, Moroccan, and Arabic) in the Flemish context. The results do not support the common myths in Flanders, which ascribe to the latter group lower levels of listening proficiency in the language of schooling. Although no significant differences across high and low-status language groups on the listening test could be found, this finding does not challenge the assumption that low-status language groups are at higher risk for school failure. Their greater academic risk may be due to other factors, such as a lack of learning opportunities or lower self-confidence.

Some suggestions should be taken into account for future use of the listening instrument. First, the 24 test items measured most accurately the lower and average ability listening levels. From this, we can infer that the comprehensive listening test was relatively easy for the target group. Future research may improve upon the developed listening test by increasing the difficulty level of the test items and enclosing new and more difficult items. Future research could also focus on other sources of evidence to contribute to the validity of the listening test. For example, external validity could be explored by comparing listeners' performance on the comprehensive listening test with their scores on other oral tests. Evidence-based on response processes can come from eye movements or think-aloud verbal protocol approaches (AERA et al., 2014). Finally, the findings of this study were collected in Flanders and the generalization of these findings should be made with caution. Future research could replicate the validation procedures for the development of a comprehensive listening test in different countries.

Despite the limitations, the results of this study can have some implications for listening test development and practice. First, IRT for dichotomously scored items is a useful approach to improve the psychometric quality of listening measures. Further, if the test scores are to be used to compare subgroups of students, it is interesting to investigate whether the test items offer a sizable advantage to a particular subgroup and the observed group difference is due to the presence of DIF (Jang & Roussos, 2009). A finding of the current study is that the test items assessing students' ability to define difficult words in the listening text may give a sizable advantage to boys or girls. Because of the importance of vocabulary knowledge for the construct validity of listening skills, it is impossible to fully eliminate the influence of vocabulary prior knowledge. In this respect, we highlight the importance of developing context-dependent vocabulary items in which students mainly have to go back to the listening text to complete the test items.

To conclude, the developed listening instrument may be a practical tool for Flemish teachers and researchers to investigate students' listening comprehension skills, which, in turn, may contribute to the nearly unexplored field of listening research in the language of instruction.

References

Acat, M. B., Demiral, H., & Kaya, M. F. (2016). Measuring listening comprehension skills of 5th grade school students with the help of web based systems. *International Journal of Instruction, 9*(1), 211–224. https://doi.org/1.12973/iji.2016.9116a

Adelmann, K. (2012). The art of listening in an educational perspective: listening reception in the mother tongue. *Education inquiry, 3*(4), 513-534.https://doi.org/10.3402/-edui.v3i4.22051

Agirdag, O. (2010). Exploring bilingualism in a monolingual school system: insights from Turkish and native students from Belgian schools. *British Journal of Sociology of Education, 31*(3), 307-321. https://doi.org/10.1080/01425691003700540

Akker, E., & Cutler, A. (2003). Prosodic cues to semantic structure in native and nonnative listening. *Bilingualism: Language and Cognition*, *6*(2), 81–96. https://doi.org/10.1017/S1366728903001056

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: Author.

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: an individual differences approach. *Language Learning,* 62(s2), 49–78. https://doi.org/1.1111/j.1467-9922.2012.00706.x

Aryadoust, V., Goh, C.C.M., & Kim, L.O. (2011) An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly, 8*(4), 361-385. https://doi.org/10.1080/15434303.2011.628632

Asia, A., Tolla, A., & Salam, S. (2019). Indonesian vocabulary mastery of early-aged children in Paud Melati Makassar. *Journal of Language Teaching and Research, 10*(3), 535-540. https://doi.org/10.17507/jltr.1003.17

Bachman, L., & Palmer, A. (2010). *Language assessment in practice.* Oxford University Press.

Beall, L. M, Rosier-Gill, J., Tate, J., & Matten, A. (2008). State of the context: listening education. *The International Journal of Listening, 22*(2), 123–132. https://doi.org/10.1080/10904010802174826

Berninger, V., & Abbott, R. (2010). Listening comprehension, oral expression, reading comprehension and written expression: Related yet unique language systems in grades 1, 3, 5, and 7. *Journal of Educational Psychology, 102*(3), 635–651.

https://doi.org/10.1037/a0019319

Blommaert, J., & Van Avermaet, P. (2008). *Taal, onderwijs en de samenleving: De kloof tussen beleid en realiteit.* [Language, education and the society: The gap between policy and reality]. Antwerp, Belgium: EPO.

Bourdieu, P. (1991). *Language and symbolic power.* Harvard University Press.

Brown, G. (2008). Selective listening. *System, 36*, 10-21. https://doi.org/10.1016/-j.system.2007.11.002

Brownell, J. (2016). *Listening: Attitudes, principles, and skills* (6th ed.). Pearson.

Buck, G. (2001). *Assessing listening.* Cambridge University Press.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/-jss.v048.i06

Chang, A. C., & Read, J. (2013). Investigating the effects of multiple-choice listening test items in the oral versus written mode on L2 listeners' performance and perceptions. *System, 41*(3), 575-586. https://doi.org/10.1016/j.system.2013.06.001

Chapelle, C. A. (1999). *Validity in language assessment. Annual review of applied linguistics, 19,* 254-272. https://doi.org/ 10.1017/S0267190599190135

Crocker L., & Algina J. (2006). *Classical and modern test theory.* Boca Raton, FL: H. B. Jovanovich.

Davidson, F. (2004). The identity of language testing. *Language Assessment Quarterly: An International Journal, 1*(1), 85-88. https://doi.org/10.1207/s15434311laq0101_9

de Ayala, R. J. (2009). *The theory and practice of item response theory*. The Guilford Press.

De Champlain, A. F. (2009). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44*(1), 109-117. https://doi.org/ 10.1111/j.1365-2923.2009.03425.x

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16,* 5-18. https://doi.org/10.1007/s11136-007-9198-0

Embretson, S. E., & Reise, S. P. (2013). *Item response theory.* Psychology Press.

Ferne, T., & Rupp, A. (2007). A synthesis of 15 years of research on DIF in language testing: methodological advances, challenges, and recommendations. *Language Assessment Quarterly: An International Journal, 4*, 113-148. https://doi.org/10.1080/15434300701375923

Field, J. (2013). *Examining listening* (pp. 77–151). Cambridge University Press.

Flowerdew, J., & Miller, L. (2010). Listening in a second language. In A.D. Wolvin (Ed.), *Listening and Human Communication in the 21st Century* (pp. 158-177). UK Blackwell.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing (19)*2, 133–167. https://doi.org/10.1191/0265532202lt225oa

Goh, M., & Aryadoust, V. (2016). Learner listening: new insights and directions from empirical studies. *International Journal of Listening, 30*(1-2), 1-7. https//doi.org/10.1080/109040-18.2016.1138689

Green, R. (2017). *Designing listening tests: a practical approach*. Palgrave Macmillan.

Hagtvet, B. E. (2003). Listening comprehension and reading comprehension in poor decoders: Evidence for the importance of syntactic and semantic skills as well as phonological kills. *Reading and Writing, 16*(6), 505-539. https://doi.org/10.1023/A:1025521722900

Haladyna, M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-333. https://doi.org/10.1207/S15324818AME1503_5

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage Publications, Inc.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (2013). *Item response theory: Principles and applications.* Sage Publications, Inc.

Hogan, T. P., Adlof, S. M., & Alonzo, C. N. (2014). On the importance of listening comprehension. *International Journal of Speech-Language Pathology, 16*(3), 199–207. https://doi.org/10.3109/17549507.2014.904441

Iwankovitsch, R. (2001). The importance of listening. *Language Arts Journal of Michigan, 17*(2), 5-6. https://doi.org/10.9707/2168-149X.1314

Jang, E. E., & Roussos, L. (2009). Integrative analytic approach to detecting and interpreting L2 vocabulary DIF. *International Journal of Testing, 9*(3), 238-259. https://doi.org/10.1080/15305050903107022

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Karimi, M. N., & Naghdivand, R. (2017). Literal and inferential listening comprehension: The role of L1 vs. L2 auditory working memory capacity. *Journal of Modern Research in English Language Studies, 4*(4), 67-84. https://doi.org/10.30479/ELT.2017.1532

Kim, Y.-S. (2015). Direct and mediated effects of language and cognitive skills on

comprehension of oral narrative texts (listening comprehension) for children. *Journal of Experimental Child Psychology, 141*, 101-120. https://doi.org/10.1016/j.jecp.2015.08.003

Lau, K.-L. (2017): Strategy use, listening problems, and motivation of high- and low-proficiency Chinese listeners. *The Journal of Educational Research, 110*(5), 503-514. https://doi.org/10.1080/00220671.2015.1134421

Lehto, J. E., & Antilla, M. (2003). Listening comprehension in primary level grades two, four and six. *Scandinavian Journal of Educational Research, 47*(2), 133–143. https://doi.org/10.1080/00313830308615

Lin, S. W., Liu, Y., Chen, S. F., Wang, J. R., & Kao, H. L. (2015). Development of a computer-based measure of listening comprehension of science talk. *International Journal of Science and Mathematics Education, 13*(6), 1469-1486. https://doi.org/10.1007/s10763-014- 9559-4

Lynn, R., & Mikk, J. (2009). Sex differences in reading achievement. *Trames, 13*(1), 3–13. https://doi.org/10.3176/tr.2009.1.01.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847-62. https://doi.org/0.3758/BRM.42.3.847

Marx, A., Heppt, B., & Henschel, S. (2017). Listening comprehension of academic and everyday language in first language and second language students. *Applied Psycholinguistics, 38*(3), 571-600. https://doi.org/10.1017/S0142716416000333

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models, *Measurement, 11*(3), 71-101. https://doi.org/10.1080/15366367.2013.831680

McKendry, M. G., & Murphy, V. A. (2011). A comparative study of listening comprehension measures in English as an additional language and native English-speaking primary school children. *Evaluation & Research in Education, 24*(1), 17–4. https://doi.org/1.1080/0950079.201.531702

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher, 18*(2), 5-11.https://doi.org/10.3102%2F0013189-X0-18002005

Messick, S. (1996). Validity and washback in language testing. *Language testing, 13*(3), 241-256. https://doi.org/ https://doi.org/10.1177/026553229601300302

Neumannn I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. *International Journal of*

Science Education, 33(10), 1373-1405. https://doi.org/10.1080/09500693.2010.511297

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing, 24*(4), 517-537. https://doi.org/10.1177/026553-2207080771

Oduolowu, E. O., & Oluwakemi, A. E. (2014). Effect of storytelling on listening skills of primary one pupil in Ibadan North Local Government area of Oyo State, Nigeria. *International Journal of Humanities and Social Science, 4*(9), 100-107.

Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2013). Analysis of sources of latent class DIF in international assessments. *International Journal of Testing, 13*(3), 272–293. https://doi.org/10.1080/15305058.2012.738266

Özbay, M. (2010). Turkish education neglected area: listening training. *Turkish Language Teaching Articles.* Ankara: Oncu Book, 191-201.

Papadopoulou, D., & Clahsen, H. (2003). Parsing strategies in LI and L2 sentence processing: A study of relative clause attachment in Greek. *Studies in Second Language Acquisition, 25*, 501–528. https://doi.org/10.1017/S0272263103000214

Piolat, A., Olive, T., & Kellogg, R. T. (2005). Cognitive effort during note taking. *Applied Cognitive Psychology, 19*(3), 291-312. https://doi.org/10.1002/acp.1086

Potocki, A., Ecalle, J., & Magnan, A. (2012). Narrative comprehension Skills in 5-Year-Old children: correlational analysis and comprehender profiles. *The Journal of Educational Research, 106*(1), 14-26. https://doi.org/10.1080/00220671.2012.667013

Pulinx, R., & Van Avermaet, P. (2014). Linguistic diversity and education. Dynamic interactions between language education policies and teachers' beliefs. A qualitative study in secondary schools in Flanders (Belgium). *Revue Française de Linguistique Appliquée, 19*(2), 9-27. https://doi.org//10.3917/rfla.192.0009

Pulinx, R., Van Avermaet, P., & Agirdag, O. (2017). Silencing linguistic diversity: The extent, the determinants and consequences of the monolingual beliefs of Flemish teachers. *International Journal of Bilingual Education and Bilingualism, 20*(5), 542-556. https://doi.org/10.3917/rfla.192.0009

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36. https://doi.org/10.1177/0146621697211002

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108-116. https://doi.org/10.7334/psicothema2013.260

Roever, C., & McNamara, T. (2006). Language testing: the social dimension. *International Journal of Applied Linguistics, 16*(2), 242-258. https://doi.org/ 10.1111/j.1473-

4192.2006.00117.x

Rost, M. (2011). *Teaching and researching listening* (2nd ed.). Pearson Education.

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24*(3), 293-322. https://doi.org/10.3102/10769986024003293

Rupp, A. A., & Leighton, J. P. (Eds.). (2016). *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications.* John Wiley & Sons.

Santos, S., Viana, F., Ribeiro, I. Prieto, G., Brandao, S., & Cadime, I. (2015). Development of listening comprehension tests with narrative and expository texts for Portuguese students. *Spanish Journal of Psychology, 18*(e5), 1-7. https://doi.org/10.1017-/sjp.2015.7

Serraj, S., & Noordin, N. (2013). Relationship among Iranian EFL students' foreign language anxiety, foreign language listening anxiety and their listening comprehension. *English Language Teaching, 6*(5), 1–12. https://doi.org/10.5539/elt.v6n5p1

Siegel, J. (2013). Second language learners' perceptions of listening strategy instruction. *Innovation in Language Learning and Teaching, 7*(1), 1–18. https://doi.org/10.1080/-17501229.17502011.17653110.

Song, X., Southern, G., & Klinger, D. (2015). DIF investigations across groups of gender and academic background in a large-scale high-stakes language test. *Papers in Language Testing and Assessment, 4*(1).

Spaan, M. (2007). Evolution of a test item. *Language Assessment Quarterly, 4*(3), 279-293. https://doi.org/10.1080/15434300701462937

Sulaiman, N., Muhammad, A. M., Ganapathy, N. N. D. F., Khairuddin, Z., & Othman, S. (2017). Students' perceptions on using different listening assessment methods: Audio-only and video media. *English Language Teaching, 10*(8), 93-99. https://doi.org/-10.5539/elt.v10n8p93

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K.A. Bollen, & J.S. Long (eds.), *Testing Structural Equation Models* (pp. 10-39). Sage.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*, 159–203. https://doi.org/10.1177/0146621603027003001

Tucker, S. (2007). Using remark statistics for test reliability and item analysis. Retrieved from https://www.umaryland.edu/media/umb/cits/umbtestscoring_testanditemanalysis.pdf

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment*

Quarterly, 5(3), 218–243. https://doi.org/10.1080/15434300802213015

Wagner, E. (2013). Assessing listening. In A. J. Kunan (Ed.), *Companion to language assessment* (Vol. 1, pp. 47–63). Wiley-Blackwell.

Weir,, C. (2005). *Language testing and validation: An evidence based approach.* Palgrave.

Wolfgramm, C., Suter, N., & Göksel, E. (2016). Examining the role of concentration, vocabulary and self-concept in listening and reading comprehension. *International Journal of Listening, 30*(1-2), 25-46. https://doi.org/10.1080/10904018.2015.1065746

Wolvin, A.D. (2012). Listening in the General Education Curriculum. *International Journal of Listening, 26*(2), 122-128. https://doi.org/10.1080/10904018.2012.678201

Yanagawa, K., & Green, A. (2008). To show or not to show: the effects of item stems and answering options on performance on a multiple-choice listening comprehension test. *System, 36,* 107-122. https://doi.org/10.1016/j.system.2007.12.003

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement, 3*, 187-213. https://doi.org/10.-1111/j.1745-3984.1993.tb00423.x

Yen, W. M., & Fitzpatrick, R. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). American Council on Education and Praeger Publishers.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Lawrence Erlbaum Associates, Publishers.

Zubairi, A.M, & Kassim. N.L.A. (2006). Classical and Rasch analysis of dichotomously scored reading comprehension test items. *Malaysian Journal of ELT Res, 2*, 1-20.

Zumbo, B. D., & Chan, E. K. (2014). *Validity and validation in social, behavioral, and health sciences (Vol. 54).* Springer International Publishing.

Table 1

Test Framework with Components, Sub-skills, and Example Items (translated from Dutch).

Component	Sub-skill	Example item
1. Literal	1.1. Defining the literal meaning of a word or a word group. 1.2. Remembering facts that have been explicitly mentioned in the text. 1.3. Identifying detailed information • persons and objects • numbers • place and time.	**Example item for sub-skill 1.1** *What is the task of an alert dog?* A. Warning his owner. B. Curing his owner. C. Helping his owner. D. Protecting his owner. **Example item for sub-skill 1.2** *What does Sammy do when Emily's sugar level is too low?* A. Sammy starts barking. B. Sammy presses the alarm button. C. Sammy starts wagging. D. Sammy warns Emily's parents.
2. Inferential	2.1. Deriving the implicit meaning of a word or a word group. 2.2. Linking different sentences or text parts • cause and effect • reason and explanation • comparison and contrast • means and ends. 2.3. Identifying the global content of the text.	**Example item for sub-skill 2.2** *Why is Sammy such a special dog?* A. He smells better than other dogs. B. He is the first dog that was trained by Laura. C. He is the first dog that can help children with diabetes. D. He is smarter than other dogs.

Note. More information about the stimulus text can be found in Appendix A.

Table 2

Item Characteristics from the Classical Item Analysis.

Item	*p-value*	Item-total correlation	Item	*p-value*	Item-total correlation
1	.901	.252	24	.690	**.208**
2	.781	**.017**	25	.876	**.082**
3	.811	**.238**	26	.837	**.145**
4	.803	.369	27	.580	.358
5	.588	**.235**	28	.600	.254
6	.895	**.218**	29	.439	.362
7	.597	.348	30	.812	.273
8	.690	.337	31	**.956**	**.135**
9	.754	.274	32	.452	.371
10	.325	.330	33	.857	.260
11	.504	.283	34	.675	**.138**
12	.847	.250	35	.629	**.234**
13	**.936**	**.069**	36	.886	.248
14	.359	**.114**	37	.529	**.148**
15	.786	.277	38	.871	.314
16	.886	**.127**	39	.627	.369
17	.534	.272	40	.843	.350
18	.811	.274	41	.817	.351
19	.895	.280	42	.846	.248
20	**.917**	**.179**	43	**.937**	**.137**
21	.651	.307	44	.678	.256
22	.780	**.200**	45	.818	.275
23	**.944**	**.161**	46	.618	**.119**

Table 3

Factor Loadings of the NLFA.

Item	One factor solution (n = 27)	One factor solution (n = 24)	Item	One factor solution (n = 27)	One factor solution (n = 24)
1	.519	.527	28	**.263**	
4	.560	.563	29	.382	.397
7	.421	.447	30	.355	.355
8	.396	.394	32	.443	.443
9	.347	.346	33	.341	.345
10	.373	.363	36	.429	.428
11	.297	.304	38	.494	.469
12	.295	.304	39	.432	.432
15	.352	.341	40	.531	.531
17	**.262**		41	.484	.471
18	.324	.316	42	.311	.301
19	.474	.470	44	**.231**	
21	.351	.355	45	.348	.355
27	.379	.376			

Note. n = number of items

Table 4

Absolute and Relative Fit Indices.

	Rasch	2PLM
SRMSR	.038	.030
MADaQ3	.028	.027
AIC	24583	24568
BIC	24706	24804

Table 5

Item Infit Statistics.

Item	1PLM Infit	1PLM Infit_*p*	2PLM Infit	2PLM Infit_*p*		Item	1PLM Infit	1PLM Infit_*p*	2PLM Infit	2PLM Infit_*p*
1	0.967	1	0.973	1		13	1.005	1	1.021	1
2	0.953	1	0.953	1		14	1.006	1	1.004	1
3	0.985	1	0.985	1		15	1.012	1	1.022	1
4	1.009	1	1.008	1		16	0.993	1	0.991	1
5	1.019	1	1.029	1		17	1.010	1	1.007	1
6	1.020	1	1.015	1		18	0.981	1	0.987	1
7	1.038	.379	1.040	.639		19	0.974	1	0.979	1
8	1.014	1	1.024	1		20	0.990	1	0.987	1
9	1.016	1	1.018	1		21	0.950	1	0.962	1
10	1.025	1	1.032	1		22	0.974	1	0.976	1
11	0.977	1	0.979	1		23	1.022	1	1.029	1
12	1.020	1	1.017	1		24	1.007	1	1.016	1

Table 6

DIF-Statistics.

Item	MHdelta gender (effect sizes)	*p*-values gender	Item	MHdelta gender (effect sizes)	*p*-values gender
1	-1.500 (B)	.023*	13	0.657	.178
2	-1.021 (B)	.018*	14	-0.272	.436
3	0.114	.905	15	-1.124 (B)	.007**
4	-1.139 (B)	.001***	16	1.751 (C)	.035*
5	-0.550	.128	17	-0.361	.469
6	0.778	.059	18	0.158	.675
7	1.731 (C)	.000***	19	0.528	.419
8	-0.545	.098	20	0.044	.936
9	0.618	.064	21	0.597	.261
10	-0.238	.494	22	0.034	.977
11	-1.251 (B)	.003**	23	-0.281	.570
12	-0.668	.271	24	0.517	.143

Note. Effect size: A: Small effect, B: Moderate effect, C: Large effect.

p-value: .01*, .001**, .000***

Table 7

Item Ability Scale: Difficulty and Discrimination Indices of the Final Items.

Item	Sub-skill	Text type	QT	Difficulty	Discrimination
6	Literal (Identifying detailed information: objects)	Informative	O.E.	-1.203	0.659
16	Inferential (Deriving the implicit meaning of a word)	Informative	M.C.	-0.277	0.768
14	Inferential (Identifying the global content of the text)	Informative	O.E.	-0.245	0.755
7	Literal (Identifying the explicit meaning of a word)	Informative	M.C.	0.121	0.586
13	Inferential (Making inferences between text parts)	Informative	M.C.	0.377	0.668
3	Inferential (Making inferences between text parts)	Informative	M.C.	0.481	0.851
20	Inferential (Making inferences between text parts)	Instruction	O.E.	0.621	0.818
12	Literal (Identifying detailed information: place)	Instruction	M.C.	0.707	0.660
4	Literal (Identifying the explicit meaning)	Informative	M.C.	1.000	0.741
5	Literal (Remembering facts)	Informative	M.C.	1.242	0.653
9	Inferential (Making inferences between text parts)	Informative	M.C.	1.453	0.693
10	Inferential (Identifying the explicit meaning)	Instruction	M.C.	1.589	0.638
24	Literal (Remembering facts)	Instruction	M.C.	1.673	0.713

22	Inferential (Identifying the global content of the text)	Instruction	M.C.	1.800	0.999
15	Literal (Remembering facts)	Informative	M.C.	1.815	0.596
23	Inferential (Identifying the global content of the text)	Instruction	M.C.	1.832	0.589
8	Inferential (Making inferences between text parts)	Informative	M.C.	1.846	0.604
17	Literal (Identifying detailed information: objects)	Informative	M.C.	1.959	0.680
2	Literal (Identifying the explicit meaning of a word)	Informative	M.C.	2.031	1.241
21	Literal (Remembering facts)	Instruction	M.C.	2.121	1.194
19	Inferential (Identifying the global content of the text)	Instruction	M.C.	2.270	1.023
18	Inferential (Making inferences between text parts)	Informative	M.C.	2.723	0.956
11	Literal (Deriving the implicit meaning of a word)	Instruction	M.C.	2.932	1.118
1	Literal (Remembering facts)	Informative	M.C.	3.301	1.025

Note. QT (Question Type): O.E. = Open-ended question, M.C. = Multiple-choice question.

Table 8

Between-group Effect for Home Language and Gender.

	F	*df*	*df2*	*p*
Home language	15.713	2	1001	.000***
Gender	1.237	1	1001	.266
Gender * home language	2.891	2	1001	.056

Note. *p*-value: .01*, .001**, .000***

Table 9

Mean Differences between the Three Language Groups.

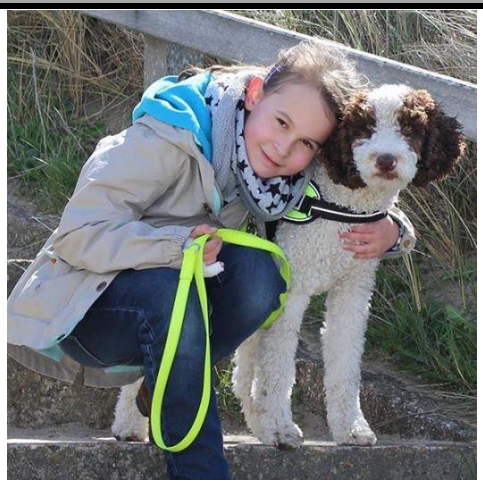
Home language	Mean difference	Std. Error	*p*
Native – High status language	.461	.137	.003**
Native – Low status language	.634	.130	.000***
High status – Low status	.173	.180	.601

Note. *p*-value: .01*, .001**, .000***

Appendix A

Script of the informative text: “A diabetes-alert dog for Emily”

A special dog



Emily has a very special dog: an alert dog called Sammy. This alert dog is trained to warn Emily when her blood sugar becomes too high or too low. Sammy watches Emily very carefully the whole time. How does Sammy know that Emily is in trouble? He smells it! He can smell very well whether the sugars in Emily's blood are normal. When the sugars are too high, Sammy smells a very sweet odor. When the sugars are too low, he will smell a very strong odor. If this is the case, he starts barking very loudly. A dog can smell 12 million times better than we do!

What is diabetes?

Diabetes is also called sugar disease. Playing, training, and going to school,... You need energy for everything you do. You can get that energy by eating, but food also contains sugar. Normally, these sugars are converted into energy in your body. This does not happen to someone with diabetes. Therefore, people with diabetes must regularly test the amount of sugar in their blood. They do this by piercing their fingers with a fine needle. A device shows them the level of sugar in their blood. If there is too much sugar in their blood, they must inject insulin, and this ensures that the sugar in their blood will be degraded. If they have too little sugar, they must eat or drink something that contains sugar.

Why does Emily need a special dog?

The sugars in Emily's blood fluctuate way too fast. Sometimes her sugar level is too high, while other times, her sugar level is too low. For most people with diabetes, getting an injection on time and paying close attention to their food can solve this problem. But for Emily this doesn't help enough. That is why her mum or dad always have to be around to take good care of her and Emily can never go to a birthday party alone or play alone with friends. Fortunately, there is Sammy. He watches Emily the whole time and warns her if something goes wrong. If there really is a big problem, Sammy can even push an alarm button! Sammy has to learn a lot. Luckily, he learns fast. But to do all this, Sammy had to practice for a full year with Laura. … Sammy is the first dog that is trained in this center and the first diabetes-alert dog.

Appendix B

Table B1

Classification of languages represented in this study.

Language	Frequency	Percentage	Status
Dutch	829	81.2	Major language
French	59	5.8	High
English	8	.8	High
German	1	.1	High
Turkish	14	1.4	Low
Berber	9	.9	Low
Polish	10	1.0	Low
Arabic	20	2.0	Low
Bulgarian	1	.1	Low
Italian	1	.1	High
Spanish	2	.2	High
Portuguese	2	.2	High
Russian	4	.4	Low
Japanese	1	.1	High
Filipino	1	.1	Low
Moroccan	1	.1	Low
Greek	4	.4	High
Danish	1	.1	High
Ukrainian	1	.1	Low
Iranian	3	.3	Low
Iraqi	1	.1	Low
Aramaic	1	.1	Low
Norwegian	1	.1	High
Indian	3	.3	Low
African	6	.6	Low
Thai	2	.2	Low
Bosnian	1	.1	Low
Armenian	1	.1	Low
American	1	.1	Low
Romanian	2	.2	Low
Congolese	1	.1	Low
Pakistani	2	.2	Low
Vietnamese	2	.2	Low
Slovaks	1	.1	Low